Technical Disclosure Commons

Defensive Publications Series

May 2020

Multimodal Hotword to Invoke a Virtual Assistant

Matthew Sharifi

Victor Carbune

Multimodal Hotword to Invoke a Virtual Assistant

ABSTRACT

To invoke a virtual assistant, a user needs to utter a pre-configured hotword. In devices that provide virtual assistant software, such utterance is detected by dedicated hardware, e.g., a low power digital signal processor (DSP). However, speaking a hotword to trigger a virtual assistant is sometimes unnatural, awkward, or inconvenient. This disclosure describes a two-stage technique that enables virtual assistant invocation without the need for the user to speak a hotword for each query. With user permission, data from multiple sensors that are integrated within a user device and/or connected to the device via fast communication protocols are utilized to detect user intent to issue a command to the virtual assistant.

KEYWORDS

- Virtual assistant
- Voice assistant
- Automatic activation
- Device activation
- Multimodal hotword
- Gaze detection
- Proximity detection
- Motion detection
- Binary classifier
- Machine learning

BACKGROUND

To invoke a virtual assistant, a user needs to utter a pre-configured hotword. In devices that provide virtual assistant software, such utterance is detected by dedicated hardware, e.g., a low power digital signal processor (DSP), that is configured to monitor ambient audio input and perform matching. However, speaking a hotword to trigger a virtual assistant may be unnatural, awkward, or inconvenient from a user experience perspective. A more seamless user experience would be, e.g., to enable the user to activate the virtual assistant by simply looking at their device or raising it and speaking as if engaged in a natural conversation.

DESCRIPTION

Many devices include multiple low power sensors. Other sensors, e.g., auxiliary cameras, etc. can be accessed via wired or wireless connections. This disclosure describes a two-stage technique that enables virtual assistant invocation without the need for the user to speak a hotword. With user permission, data from multiple sensors that are integrated within a user device and/or connected to the device via fast communication protocols are utilized to detect user intent to issue a command to the virtual assistant.

When a user intent to invoke a virtual assistant is detected, the user device is woken up from a low-power state and the virtual assistant is triggered to serve the user's request. Since data obtained from a diversity of sensors are used in determination of the user's intent to interact with a virtual assistant, the technique is referred to herein as multimodal hotword detection, where data for different modalities is obtained from different sensors.

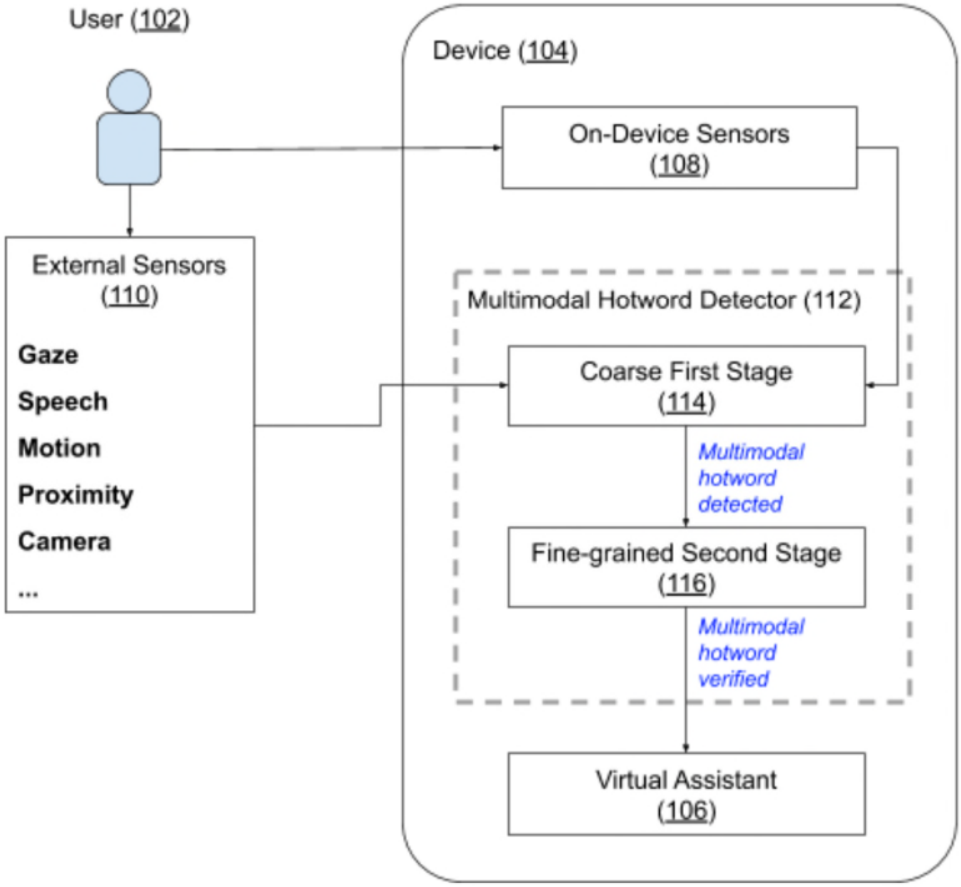

Fig. 1: Using multiple sensors to detect user intent to invoke a virtual assistant

Fig. 1 shows an example of operational implementation of the described techniques. A user (102) wishes to interact with a voice assistant (106) provided via a user device (104). The user's intent is indicated by one or more natural interactive actions, such as gaze (looking at the device), speech (starting to speak while looking towards the device), motion (raising hand that holds the device, pointing to the device), proximity to the device (walking, turning, or bending towards the device), etc.

With user permission, these natural actions are detected by various on-device sensors and external sensors (110) that operate at low power. The sensor data is processed by a two-stage multimodal hotword detector (112). The first stage is a low-power coarse stage (114) that is used

to detect likely intention to invoke the virtual assistant. The second stage, which follows the first stage, is a high-power fine-grained stage (116) to verify the intention based on combining the information obtained from various sensors. If the multimodal hotword detector determines that the user wishes to interact with the virtual assistant, the virtual assistant is invoked (which is an outcome of natural user actions) without the need for the user to utter a specific hotword.

For instance, the user can invoke a virtual assistant in a natural manner simply by raising the device and/or looking at the device and starting to talk to the virtual assistant. Detection of such interactions is achieved by combining several relevant sensor outputs, such as device orientation, device movement, user's gaze, user's speech, etc. Although the information obtained from each individual sensor may provide only a weak indication of the user's intent to invoke a virtual assistant, their combination can serve as a strong indicator of such intent.

In the coarse first stage, the various types of sensors operating at low power detect corresponding user actions that are part of a multimodal hotword. The output of the sensors operating within the first stage is accumulated into buffers. A fine-grained second stage is employed to verify the one or more candidate hotwords in the first stage output buffers. If the output of the second stage indicates that the user issued a multimodal hotword, the virtual assistant is invoked automatically, without the user specifically uttering the hotword.

If the user permits, the coarse first stage can be implemented with one or more low-power sensors within a user device, such as:

- Camera to determine the user's gaze
- Microphone to capture speech
- Accelerometer to detect device motion and orientation
- Radar sensor to detect user movements and proximity

The data from each sensor can be processed with a corresponding DSP dedicated to the sensor, such as an audio DSP for handling speech input obtained via the microphone. Alternatively, a single DSP can process the data obtained from multiple sensors. Each sensor within the first stage is associated with a corresponding detector that employs a binary classifier to indicate whether the sensor data likely contains a multimodal hotword that can be verified by the fine-grained second stage. For example, the gaze detector can determine whether the user is looking at the device, the motion detector can detect if the user has picked up the device, the radar detector can indicate if the user has placed the device close to the face, the audio detector can capture whether the user has started speaking, etc.

If one or more of the binary classifiers within the coarse first stage detect that the user likely wishes to invoke the virtual assistant, the main processor of the device is woken up from low-power operation and shifted into the fine-grained second stage (that has higher power requirements). With user permission, the buffers of sensor data accumulated by the coarse first stage serve as input to the second stage.

The second stage is implemented as a binary classifier that operates on a joint representation computed from the separate data sequences for each individual sensor within the coarse first stage. To that end, the separate data sequences can be fused to produce a joint score by employing any suitably trained machine learning model. For instance, the joint score can be generated via a deep convolutional neural network with pre-trained modules for each data stream whose outputs are subsequently processed by a binary classifier based on Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), or Quasi-Recurrent Neural Network (QRNN), etc. For example, Conv-1D blocks can be pre-trained for microphone inputs, Conv-2D blocks pre-trained for image frames and radar sensors, etc. The final output of the binary

classifier verifies that the joint representation of the sensor data obtained from the coarse first stage contains a multimodal hotword. Upon such verification, the virtual assistant is triggered.

If the user permits, the fine-grained second stage can be implemented such that multiple variants of the model can be operationalized as necessary. For instance, the output of the second stage classifier can be based on the data of a subset of the sensors. Such flexibility can help support situations when data from specific sensors is unavailable or unusable. Alternatively, or in addition, the second stage can incorporate data from additional sensors that was not included in the coarse first stage, e.g., due to the inability to operate continuously on low power. For instance, in addition to accelerometer and microphone data used in the coarse first stage, the fine-grained second stage can include data captured via sensors that require higher power consumption, e.g., a camera to capture images.

The techniques described herein can be implemented in a robust manner such that it is easy to add or remove sensors used for detecting multimodal hotwords, without substantially affecting hotword detection. Further, if the user permits, the techniques can obtain and utilize data from sensors that are external to the device, e.g., an external camera, motion sensor, etc. The threshold values used by the various models and classifiers can be set by the developers and/or specified by the user and/or configured dynamically at runtime.

Implementation of the techniques described in this disclosure with permission enables users to initiate interactions with a virtual assistant in a natural manner, thus improving the UX of devices with virtual assistant functionality. The techniques can be implemented within any devices with virtual assistant capabilities, such as smartphones, smart speakers/displays, smart appliances, wearables, Internet of Things (IoT) devices, etc.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information obtained from one or more sensors of a user device, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes a two-stage technique that enables virtual assistant invocation without the need for the user to speak a hotword. With user permission, data from multiple sensors that are integrated within a user device and/or connected to the device via fast communication protocols are utilized to detect user intent to issue a command to the virtual assistant. A coarse first stage of detection utilizes data from various types of sensors to detect the user's intention to invoke a virtual assistant. A fine-grained second stage verifies the candidate multimodal hotword (data from various sensors) in the fused sensor data. If the output of the second stage indicates that the user issued a multimodal hotword, the virtual assistant is invoked to detect and respond to the user. Use of a multimodal hotword enables users to initiate interactions with a virtual assistant in a natural manner and provides an improved user experience.

REFERENCES

1. Burke, Dave, Michael J. Lebeau, Konrad Gianno, Trausti Kristjansson, John Nicholas Jitkoff, and Andrew W. Senior. "Multisensory speech detection." U.S. Patent 9,009,053, issued April 14, 2015.
2. Mixter, Kenneth, Yuan, Yuan, and Nguyen, Tuan. "Adapting automated assistant based on detected mouth movement and/or gaze." European Patent Application EP3596584A1, published January 22, 2020.